Supplemental material: Geometry Fidelity for Spherical Images

Anders Christensen[†,2,3], Nooshin Mojab[1], Khushman Patel[1],
Karan Ahuja[1,4], Zeynep Akata[3,5,6], Ole Winther[2,7,8],
Mar Gonzalez-Franco[1], and Andrea Colaco[1]

[1] Google, USA
[2] Technical University of Denmark, Denmark
[3] Helmholtz Munich, Germany
[4] Northwestern University, USA
[5] Technical University of Munich, Germany
[6] Munich Center of Machine Learning, Germany
[7] University of Copenhagen, Denmark
[8] Copenhagen University Hospital, Denmark
andchri@dtu.dk, nooshinmojab@google.com

1 Qualitative evaluation of OmniFID

To qualitatively evaluate our proposed Omnidirectional FID, we compute FID and OmniFID on generated images from three different checkpoints of a finetuned text-to-image generative model. The model is based on a version of Imagen [3] trained on internal datasources, and finetuned using Dreambooth [2] with a batch size of 16. We finetune the model on the 360-Indoor equirectangular image dataset [1] and use captions generated by a multimodal language model. This gives us 3252 image-caption pairs after removing duplicate and empty captions.

The captions were generated by giving the multimodal model prompts with few-shot examples describing the content of corresponding equirectangular images, followed by keywords of e.g. style, lighting, and indoor/outdoor. An example of such a given few-shot example caption is: "living room with couches, TV, coffee tables and fireplace. french style decoration, daylight, indoor". Finally, we edit the prompt to be "a panoramic view of a <caption>".

Below, we show example equirectangular image generations from model checkpoints after 5000, 10000, and 20000 steps. The visualized generations are generated from the same prompts across the different checkpoints, where the corresponding prompts were selected randomly. Results show that the FID score is near-constant across the checkpoints (33.96, 35.42, 34.95, respectively). Further, although the example generations from the 5000 step model demonstrate that the model has issues constructing realistic geometry, the FID score is lowest for this checkpoint. On the contrary, OmniFID decreases monotonically over the checkpoints as geometry fidelity improves (63.39, 60.38, 55.07, respectively).

[†] Work done at Google, USA

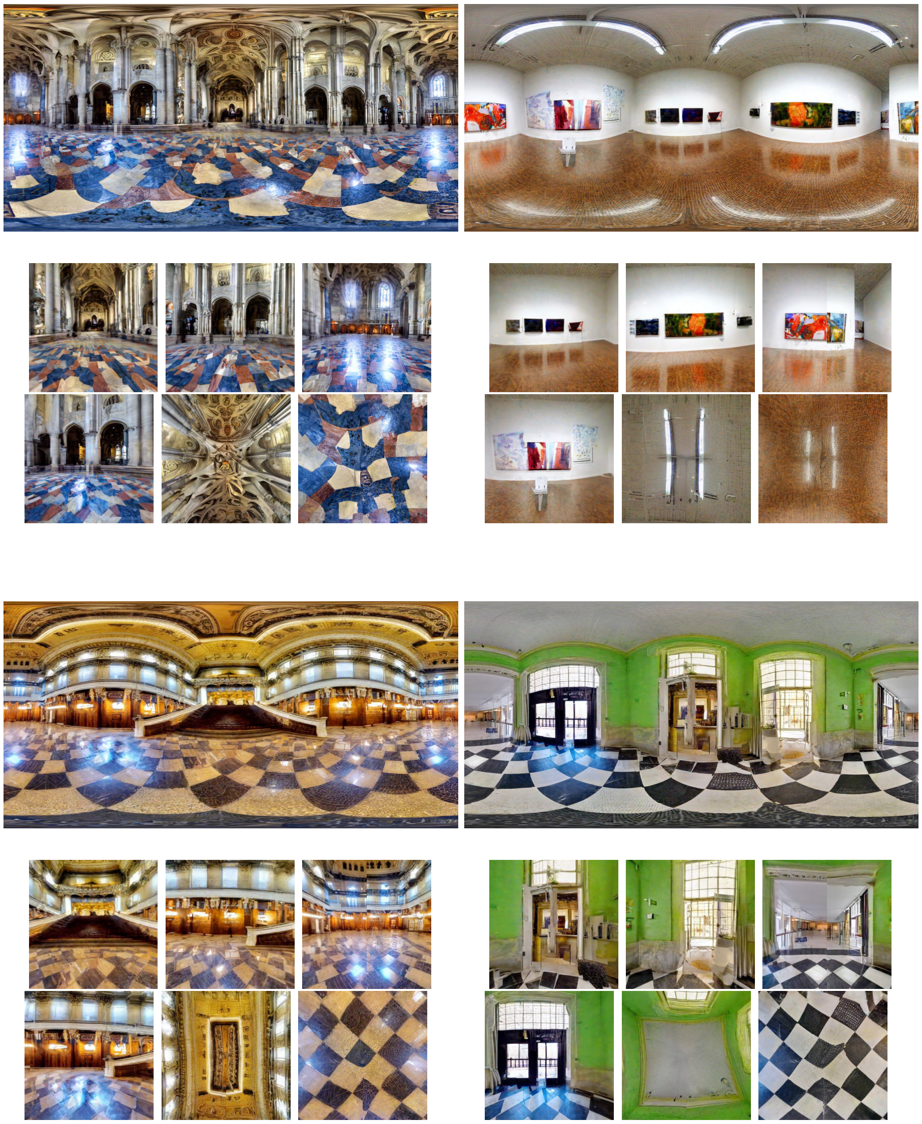

Fig. 1: Four example equirectangular generations of a text-to-image model fine-tuned on 360-Indoor after 5000 steps. Under each generation we show the cubemap images to illustrate the geometry of the rendered views (top left to bottom right: front/right/back/left/up/down). FID is 33.96, OmniFID is 63.39.

.

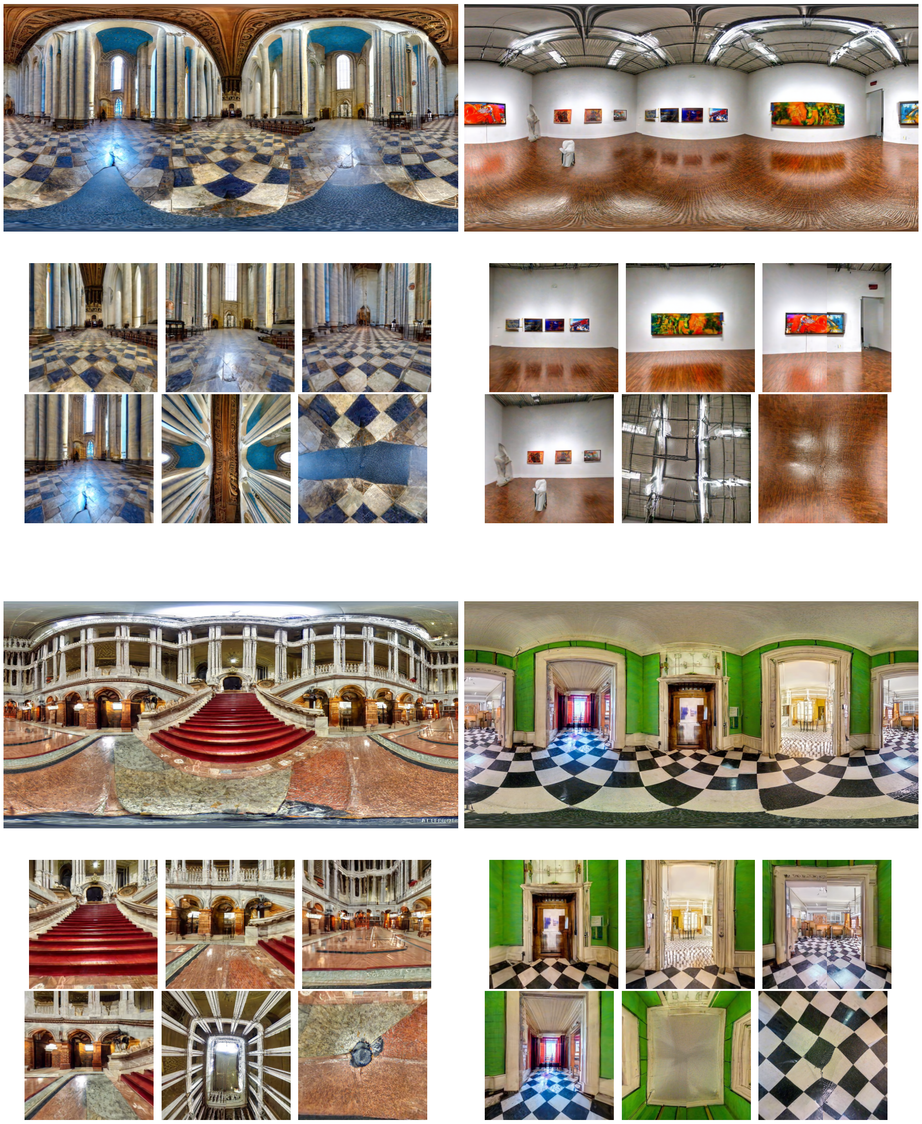

Fig. 2: Four example equirectangular generations of a text-to-image model fine-tuned on 360-Indoor after 10000 steps. Under each generation we show the cubemap images to illustrate the geometry of the rendered views (top left to bottom right: front/right/back/left/up/down). FID is 35.42, OmniFID is 60.38.

.

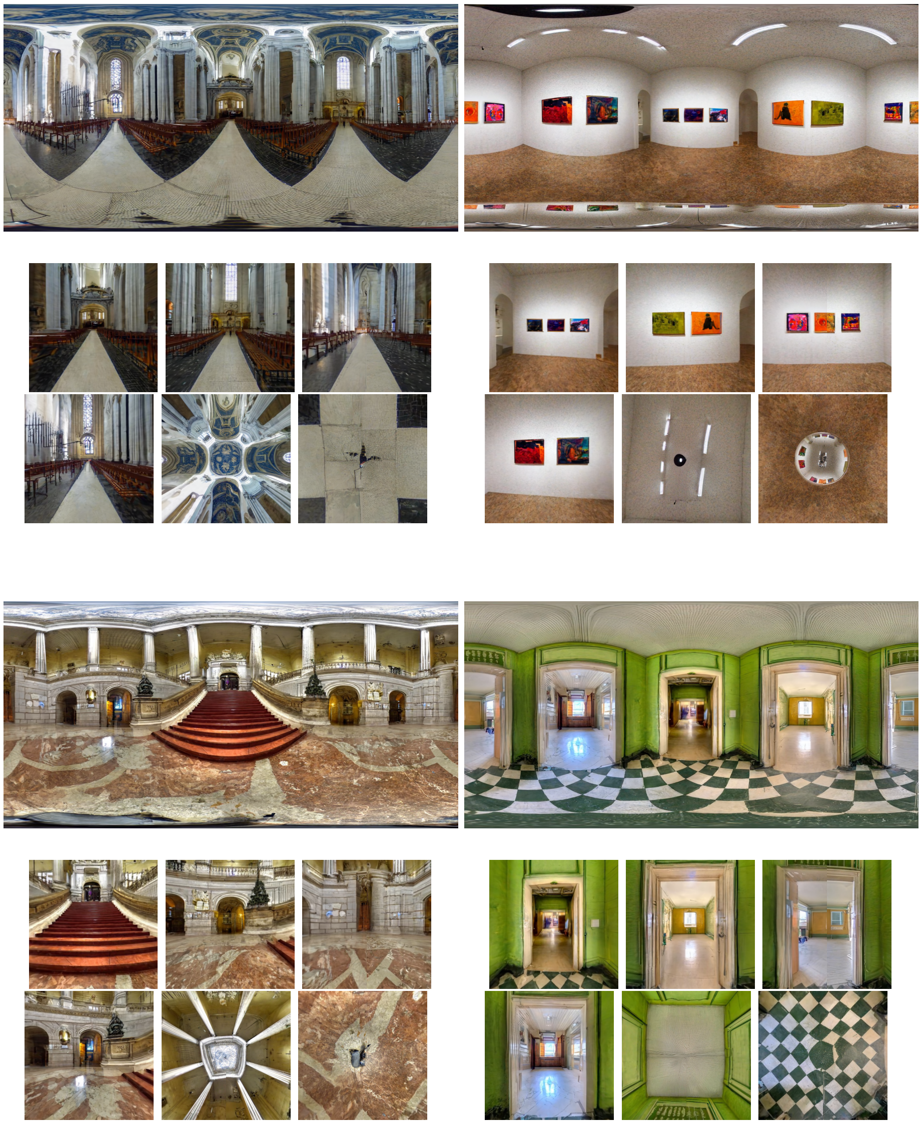

Fig. 3: Four example equirectangular generations of a text-to-image model fine-tuned on 360-Indoor after 20000 steps. Under each generation we show the cubemap images to illustrate the geometry of the rendered views (top left to bottom right: front/right/back/left/up/down). FID is 34.95, OmniFID is 55.07.

.

References

1. Chou, S.H., Sun, C., Chang, W.Y., Hsu, W.T., Sun, M., Fu, J.: 360-indoor: Towards learning real-world objects in 360 indoor equirectangular images. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 834–842 (2019)
2. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
3. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)